Memoria Proyecto Final

Álvaro Beltrán y Yábir García

June 8, 2020

Índice

1 Definición del problema a resolver y enfoque elegido.

2 Argumentos a favor de la elección de los modelos.

En cuanto a los modelos lineales elegidos nos hemos decantado por Regresión Logística y Perceptron, debido a que son dos modelos que hemos estudiado en clase para ejemplos de clasificación binaria, como es nuestro caso.

Además parece interesante probar con el Perceptron puesto que también vamos a usar el perceptron multicapa (MLP). Y puesto que Regresión Logística ha dado tan buenos resultados en las prácticas de esta asignatura y suele funcionar muy bien en clasificación parece indispensable probar este modelo.

3

4 Valoración del interés de la variables para el problema y selección de un subconjunto.

En principio, tras leer la descripción de los atributos estudiados. Solo vemos dos variables que no aportan información al problema. La primera es "Education-num" que es una representación numérica del atributo education, no aporta nada pues hemos tomado la decisión de dividir cada variable categórica en una serie de variables donde asignamos 1 si es de un determinado tipo del dominio de la variable o 0 si no es de ese tipo. La segunda variable es "fnlwgt" que determina el número de personas representadas con esa instancia, entendemos que esa variable no aporta información de fuera de la muestra y por lo tanto no es interesante.

5

6 Justificación de la función de pérdida usada.

Para este problema hemos elegido exactitud (Accuracy) por que estamos buscando como objetivo no equivocarnos en la clasificación.

$$\frac{TP + TN}{TP + TN + FP + FN}$$

Además, como se ve en la fórmula es muy intuitivo y simple, factor que nos ayudará a la hora interpretar los resultados.

7

8 Estimación de los hiperparámetros.

Para los modelos estudiados anteriormente hemos seleccionado una serie de parámetros y vamos a hacer inferencia sobre otros:

- **Regresión Logística**. Hemos decidido usar el solver lbfgs que es newton pero con mejoras en memoria para reducir el tiempo, vamos a usar este porque newton adapta la tasa de aprendizaje según le convenga y por tanto es más eficaz. Usamos penalización L1 ya que usamos Lasso. Y hacemos inferencia sobre los parámetros C y tol. El parámetro C es el coeficiente que acompaña a la penalización y tol es la tolerancia usada en el modelo.

- **Perceptron**. Hemos prefijado la máximas iteraciones a 2000 por que probando con menos no acababa y hemos decidido que baraje los datos en cada iteración del perceptron como hemos visto en teoría. Hacemos inferencia sobre los parámetros alpha y tol. El parámetro alpha es el coeficiente que acompaña a la penalización (elegida la que viene por defecto) y tol es la tolerancia usada en el modelo.

- **Ramdom Forest**. Vamos a usar bootstrap porque es capaz de medir la incertidumbre de nuestro modelo mediante una técnica de reelección de muestras. Para determinar las máximas características del árbol vamos a usar la raíz cuadrada por que hay evidencias empíricas de que es el mejor. Hemos hecho inferencia sobre el criterio de selección, sobre si elegir entropy o gini.